CCPROG2 MP PART 1 [40% of the MP] *サルバドール・フロランテ*
AY2023 Term 2

GENERAL INSTRUCTIONS

1. Honesty Policy and Honor Code apply. Please do not cheat. **The consequence of cheating is a grade of 0.0 (fail) for the course.**
2. Make sure that you have read and understood the problem requirements.
3. Comply with the specifications and restrictions stated in MP Specs document, and inside the comments written in the skeleton codes. Non-compliance will result in deductions.
4. **You are NOT allowed to use pre-defined library functions that we did not discuss in class (unless specified otherwise).**
5. **Do NOT use float data type. Always use double data type for floating point numbers in all MP challenges.**
6. Subject your solution to exhaustive testing to ensure that the solution is logically correct. The following scoring and deductions system will be applied.
    - A perfect score will be awarded to a compliant and logically correct solution.
    - Deductions will be applied based on the severity of the logical error. In the worst case, scenario, a score of 0 will be given (meaning the solution is logically incorrect).
    - Each unique compiler warning will result in a deduction of one point. Do not forget to use -Wall compiler option.
    - A syntax error will result in a score of 0 for the associated challenge.
7. Submit the required deliverables before the specified Canvas deadline.
8. Question? Please post your question in our Canvas MP Discussion Thread. Note: I will not answer an MP related question sent via email unless it is personal in nature (for example, you and your partner would like to part ways because of disagreements in working habits).

Asking Real-Life Questions and Formulating C Functions as Answers Based on a Health-Related Dataset

INTRODUCTION

"Big Data" and "*Data Science"* are recent buzz words in multiple disciplines including the sciences and engineering. For this machine problem (MP), we will represent, and process real-life data from the Stage of Global Air (SoGA) website https://www.stateofglobalair.org/

Concepts covered in CCPROG2, i.e., arrays, string, structures, and file processing will be applied. Through this MP, you'll need to demonstrate that you can:

- design and implement your own data structure for representing, storing, accessing, and manipulating a given dataset
- design and implement your own algorithms as answers to real-life questions based on the dataset
- specify test cases
- test and debug programs
- properly document and articulate your solution to the MP

The MP will be accomplished in several parts, with this document describing Part 1 activities.

ACTIVITY #1: Familiarize yourself with the "Loss Life Expectancy" SoGA dataset[1].

All students must accomplish this first activity independently on his/her own.

1. Download the SoGA zip file *"Loss in life expectancy due to air pollution in 2019"* from https://www.stateofglobalair.org/resources?resource_category=data#block-exposedformresources-all
2. Read also about the highlights of the negative effects of air pollution in https://www.stateofglobalair.org/health/life-expectancy https://www.stateofglobalair.org/health/global
3. For details of the negative effects of air pollution, download & read the SoGA 2020 Report from https://www.stateofglobalair.org/sites/default/files/documents/2022-03/soga-life-expectancy_0.pdf *NOTE: You are not required to read and understand everything in the report.*
4. Check out also the other related links on your own, for example, https://www.stateofglobalair.org/air

[1] Health Effects Institute. 2022. How Does Air Pollution Affect Life Expectancy Around the World? A State of Global Air

The SoGA dataset include statistics on Baseline Life Expectancy (column E) and risk factors that cause reduction in life expectancy starting from column *F* on Air Pollution to column *S* on Unsafe Sex. Listed below are links that may help you gain an initial understanding of the nature of the risk factors and their effects on our health.

- Life Expectancy https://ourworldindata.org/life-expectancy
- Air Pollution https://www.who.int/news-room/spotlight/how-air-pollution-is-destroying-our-health
- Ambient PM2.5 https://www.epa.gov/pm-pollution/particulate-matter-pm-basics
- Ozone https://www.epa.gov/ground-level-ozone-pollution/ground-level-ozone-basics
- Household Air Pollution (HAP) https://www.who.int/news-room/fact-sheets/detail/household-air-pollution-and-health
- Environmental Hazards https://www.cdc.gov/nceh/tracking/tracking-intro.html#:~:text=EXAMPLES%20OF%20ENVIRONMENTAL%20HAZARDS%20INCLUDE%3A%201%20Air%20contaminants,consumer%20products%208%20Extreme%20temperatures%20and%20weather%20events
- Occupational Hazards https://www.webmd.com/a-to-z-guides/occupational-hazards
- Unsafe Water, Sanitation and Hygiene (WaSH) https://www.who.int/teams/environment-climate-change-and-health/water-sanitation-and-health/burden-of-disease#:~:text=Unsafe%20WASH%20is%20linked%20to%20many%20adverse%20health,in%20drinking-water%2C%20as%20well%20as%20impacts%20on%20well-being.
- Metabolic Syndrome https://www.mayoclinic.org/diseases-conditions/metabolic-syndrome/symptoms-causes/syc-20351916#Symptoms
- Dietary https://www.nih.gov/news-events/nih-research-matters/how-dietary-factors-influence-disease-risk
- High Fasting Plasma Sugar https://www.verywellhealth.com/understanding-the-fasting-plasma-glucose-test-1087680
- Tobbacco, Smoking & 2nd Hand Smoke https://www.cdc.gov/tobacco/basic_information/health_effects/index.htm
- Unsafe Sex https://www.nature.com/articles/s41598-023-40001-2

NOTE #1: you are not required to understand everything written in the links above. They are provided as supporting references for you to have at least an initial understanding of the risk factors (like WaSH and metabolic) used in the dataset.

NOTE #2: A subset of the original data will be used in our MP. In particular, except for the row containing the "Global" statistics, all rows with IsTerritory == 0 (i.e., regions such as continents) were excluded. The three columns LocationID, IsTerritory and IsRegion were also excluded. To simplify string processing, an underscore was added to a Territory/country name when the name is made up of more than one word (for example, "New Zealand" was edited as "New_Zealand"). Refer to the accompanying ***SoGA_DATASET.TXT*** *file. The 1st column is for the names of the territories, the 2nd column is for the baseline life expectancy (LE), the 3rd to the last columns are for the reductions or loss in life expectancy due to different risk factors.*

ACTIVITY #2: Choose a partner or choose to work independently. Sign-up in Canvas for the MP grouping.

Find a partner who will HONESTLY COOPERATE and COLLABORATE with you in accomplishing the MP. Partners may come from two different sections among S11A, S11B, & S12A. Alternatively, you may choose to work independently on your own.

For those who chose to work with a partner, make sure that you accomplish ALL activities specified below TOGETHER. Ideally, both partners will contribute equally to the task at hand.

ACTIVITY #3: Formulate real-life questions.

1. (Together with your partner) Come up with FIVE real-life specific questions that can be answered based on the SoGA dataset. There are **three** requirements in forming your questions:
    a. The answer to each question should involve one or a combination of the following algorithms.
        - minimum or maximum
        - count or sum or average
        - linear search
        - binary search (at least one question should be answered based on binary search)
        - selection sort (at least one question should be answered based on selection sort)
    b. At least THREE of the questions must have one or more parameters. A parameter can be a single number, a string (i.e., a territory name), or an array of numbers or an array of strings.
    c. At least ONE of the questions should produce an array of strings, i.e., a list of territory names, as answer.

Here are some example questions[2] (note: you cannot use the same exact question). Questions with parameters are those shown with angled brackets, i.e., *<parameter>.*

Q1: Which territory has the lowest base line life expectancy? Display the name of the territory and the corresponding life expectancy. [minimum]
Q2: What are the statistics for *<parameter_territory_name>*? List the values in the same order as the columns. [linear search or binary search after sorting]
Q3: What is the average loss of life expectancy due to tobacco? [average]
Q4: How many territories have a reduction of at least *<parameter_number>* year(s) in life expectancy due to air pollution? [count]
Q5: Which are the top *<parameter_number>* territories with the lowest reduction in life expectancy due to Metabolic factor? List the names of the territories with the corresponding life expectancy reduction starting from the lowest reduction value. [selection sort]

NOTE: Do NOT form questions based on region data (for example continents). Refer back to NOTE #2 in ACTIVITY #1.

2. Once you have formulated your questions, encode the following in the accompanying ***LASTNAME1_LASTNAME2.TXT*** file.
   a. your name and section [NAME1, SECTION1]
   b. the name and section of your partner [NAME2, SECTION2],
   c. your five questions [Q1, Q2, Q3, Q4, Q5]; indicate also which algorithm(s) will be used to answer for each question
   d. the expected answers to the questions [A1, A2, A3, A4, A5].
3. Rename the text file with your own last names in alphabetical order. For example, if the last name of the first student is TAN, and the last name of the second student is CRUZ, then the file should be renamed as CRUZ_TAN.TXT. If you will work independently (i.e., no partner), then the filename should correspond to your last name only, for example, SANTOS.TXT.
4. Submit the text file via the Canvas submission page for this purpose. All students must submit regardless of whether the student is working independently or with another student.

ACTIVITY #4: Answer the questions in the form of C functions.

1. First you must represent/store the data values (from the SoGA dataset) using array data structure. It is up to you to decide on (a) how many arrays you need, (b) the array sizes, (c) their dimensions and the (d) element data type. Initialize the array contents with data values read from a text file via input redirection[3]. It is up to you how to represent and store the data using arrays. Do NOT use yet the **struct** data type at this stage of the MP.
2. Answer your five questions by formulating the algorithms and implementing them as C functions. You are provided a skeleton C source file ***LASTNAME1_LASTNAME2.c*** which contains comments describing specific guidelines on what to encode in the source file.
   - If the answer to a question is just one value (for example, maximum), the C function must return the value.
   - If the answer to a question involves several values (for example the names of the top 10 countries in sample question Q5 above), then the answer should be stored in an array or arrays that will be accessed indirectly inside the C function definition.
   - The main() function should call the appropriate C function, and then call the printf() statement to display the answer to the question. Numeric answers with double data type must be displayed with 6 digits after the decimal point.
   - There should NOT be any printf() and/or scanf() statement in any function definition except in main(), and in the function that reads the SoGA data text file.
   - You must use a C **double** data type (NOT float) for all floating point values/variables/parameters/functions.
3. Make sure that the answers are NOT HARDCODED inside the function definitions.
4. Submit your source code and sample program output via the Canvas submission page for this purpose.

ACTIVITY #5: Demo your MP.

We will schedule a date/time for you to demo your project to check if it is working or not. You should know how to compile and run your program in the command line. You and your partner must both be present. The demo will be done in person *(but may be set to online mode in case of time/venue/resource related issue)*.

--- End of MP Part 1 ---

[2] Refer to the last page to see the answers to questions Q1 to Q5.
[3] Refer to the supplementary material on I/O redirection found in your Canvas home page under Module 0 on Supplementary Resources.

Preview of the MP parts 2 and 3…

MP Part 2 [40% of the MP]: expect to apply what you learned on **struct** data type (covered in Chapter 3). You will also write C functions as answers to questions. However, the questions you'll answer are those that were asked by other students. For example, the 1st group will answer the questions asked by the 8th group, while the 8th group will answer the questions asked by the 15th group, and so on… The questions will be assigned via randomization. Communication in any form with the original group who gave the questions is STRICTLY FORBIDDEN. Violation of this directive will be considered as cheating resulting to a fail grade of 0.0.

MP Part 3 [20% of the MP]: expect to apply what you learned on **file processing** (covered in Chapter 4). You have the option to choose which set of questions to answer, i.e., choose either your own five questions, or those from another group that you answered in MP Part 2.

Answers to the sample questions:

A1: Lesotho 51.727436

A2: When *parameter_territory_name* is set to "Philippines", then the answer is as follows.
Philippines
71.798423
1.675681
0.691634
0.002370
0.929484
2.383616
0.224457
0.242926
6.165008
2.162167
1.885343
2.575556
2.159935
0.460448
0.164648

A3: 1.706265

A4: When *parameter_number* is set to 1.5 years, the answer is as follows.
70

A5: When *parameter_number* is set to 10 territories, the answer is as follows.

1. Japan	2.546600
2. Lesotho	2.626596
3. Central_African_Republic	2.825729
4. France	2.991462
5. Somalia	2.991605
6. Ethiopia	2.991629
7. Chad	3.012220
8. Netherlands	3.036495
9. South_Sudan	3.067292
10. Kenya	3.074878